# Methods for estimating infant thresholds[a)]

Lynne A. Werner[b)] and G. Cameron Marean
*University of Washington, Seattle, Washington 98195*

(Received 19 July 1990; revised 6 February 1991; accepted 18 June 1991)

Detection thresholds for 1000-Hz, 16-ms tone bursts were estimated for 3- and 6-month-old infants and for young adults. The test procedure used was the observer-based psychoacoustic procedure. Thresholds were estimated using two different adaptive procedures and the method of constant stimuli. There was little difference among the average thresholds determined by any of these techniques. The least variable thresholds were obtained in the method of constant stimuli. In addition, 10 infants at each age completed two 30-trial blocks of trials in the method of constant stimuli; 22 adults completed 8 blocks of 30 trials. For 3- and 6-month-olds and for adults, there was no significant change in average threshold between blocks. Individual 6-month-olds' thresholds rarely changed by more than 5 dB between blocks, and the correlation between the thresholds obtained in the two blocks was significant. Individual 3-month-olds' thresholds, however, sometimes changed by as much as 10 dB between blocks, and the correlation between first and second block threshold was not significant. The effects of response bias on threshold were assessed by examining receiver operating characteristic plots of hit and false alarm rates at threshold and the correlation between false alarm rate and threshold. Although there was some variability in response bias, infant/observer teams tended to respond in an unbiased fashion. In one of the adaptive procedures only, false alarm rate was significantly correlated with threshold. In all procedures, the exclusion of infants with high false alarm rates changed the average thresholds obtained by less than 5 dB. About half of the infants with high false alarm rates appeared to perform no better than would be expected by chance. Finally, the difference between infant and adult thresholds for these short duration stimuli was about 10 dB greater than the difference previously reported for long duration stimuli.

PACS numbers: 43.66.Cb, 43.66.Yw [LDB]

## I. METHODS FOR ESTIMATING INFANT THRESHOLDS

In a previous paper, Olsho *et al.* (1988) reported absolute thresholds for 3-, 6-, and 12-month-old infants for 500-ms tone bursts. Thresholds of 3-month olds were 20–30 dB higher than those of adults; thresholds of 6- and 12-month olds were 10–20 dB higher than those of adults. Using the same adaptive procedure, we attempted to measure infant detection thresholds for 16-ms tone bursts in quiet. These thresholds were difficult to obtain; infants did not appear to be responsive to the short-duration tones. Moreover, the initial results were quite variable. These preliminary observations were troubling, in that we had no published work with which to compare the thresholds we were obtaining, and thus, had little basis for deciding whether the results were reasonable.

In adaptive procedures, signal level on each trial depends on the listener's performance on previous trials. Such procedures are considered an efficient means of estimating threshold. It appears that adaptive procedures would not be optimal under certain conditions, however. It is well known that infants' responses to tones and noises habituate rapidly, even when the responses are reinforced, and that habituation occurs more rapidly when the stimuli are suprathreshold (Moore *et al.*, 1975; Gray, 1987). The typical adaptive procedure used with infants begins with the stimulus at a suprathreshold level, and, on subsequent trials, the level is decreased until the infant misses a signal. If the infant's response habituates before the sensory threshold is reached, there is a concern that an adaptive procedure may track the infant's attention rather than sensitivity to the stimulus. This problem would be exacerbated in the case where the initial level of responsiveness is poor and where the rate of habituation is more rapid, as might be the case for short-duration tones.

At the same time, one would like to obtain as many trials as possible from a subject. Teller (McKee *et al.*, 1985; Teller, 1985) has made the point that the number of trials obtained from an infant in a psychophysical procedure is a major constraint on the statistical accuracy that can be achieved in the threshold estimate, whether an adaptive technique or the method of constant stimuli is used. McKee *et al.* (1985) have also shown that at least 60–100 trials may be needed to get a stable estimate of the psychometric function. While a large number of trials would be desirable from a statistical

[a)] Preliminary reports of these data were made at the 116th Meeting of the Acoustical Society of America, November, 1988 [L. Werner Olsho and G. C. Marean, J. Acoust. Soc. Am. Suppl. 1 **84**, S144 (1988)], and at the Midwinter Research Meeting of the Association for Research in Otolaryngology, February, 1990.

[b)] Requests for reprints should be addressed to this author at CDMRC, WJ-10, Box 47, University of Washington, Seattle, WA 98195.

  

point of view, then, habituation to the stimulus over trials might have the undesirable effect of reducing the slope of the psychometric function and increasing the variability of the threshold estimate.

Another important point is that we use a one-interval, yes–no procedure to test infants. In such a procedure, responsiveness, or response bias, can affect the threshold obtained whether or not the response habituates over the course of a session.

The current study, then, posed three questions. First, how do infant thresholds obtained using an adaptive procedure compare to those obtained using the method of constant stimuli? This comparison was meant to address the potential effects of habituation on threshold and threshold variability. Second, what are the effects of increasing the number of trials on infant threshold? If habituation occurs, it could have several effects. Increasing the number of trials might increase the variability in threshold estimates; threshold estimates based on performance early in a session might be better than those based on performance later in the session; and within-session reliability of threshold estimates would not be high. Third, how does response bias affect the estimate of infant threshold? In other words, do infants and observers adopt reasonable and consistent response criteria, and is response criterion related to the threshold obtained?

## II. METHOD

### A. Stimuli and apparatus

For all measures, the stimulus was a pure tone at 1 kHz. The tone burst presented to the subject had 16-ms rise/fall and no steady-state duration. We refer to this stimulus as a 16-ms tone burst. The tone was generated digitally, but the rise/fall was shaped by electronic switches. The experiment was controlled by an AT&T PC 6300 microprocessor.

Signal trials throughout the experiment consisted of 20 repetitions of the tone burst with interstimulus interval of 444 ms. No-signal trials were periods of the same total duration (9.5 s) during which no sound was presented.

Stimuli were presented to the listener's right ear using an Etymotic ER-1 insert earphone. Foam eartips, trimmed to fit various ear canal sizes, were used to hold the probe tube in place. The probe tube was also taped to the infant's pinna to keep it from being dislodged during the session. The stimulus was calibrated in a 2-cc coupler, using a Bruel & Kjaer 2215 sound level meter and a 3521A Hewlett Packard signal analyzer.

### B. Procedure

The test method was the observer-based psychoacoustic procedure (OPP) (Olsho *et al.*, 1987b). Since the method has been described in detail in previous reports (Olsho *et al.*, 1988; Olsho *et al.*, 1987a), it is described only briefly here.

The central premise of the method is that if an observer can reliably judge whether a signal or no-signal trial has been presented to an infant, when the only information upon which to base the judgment is the infant's behavior, then the infant must be responding to signals. During a session, the infant is typically seated on a parent's lap in the test booth. The infant's attention is loosely maintained ahead and at midline by an assistant manipulating toys. Neither the parent nor the assistant can hear the sounds presented to the infant. The observer watches the infant from an adjacent control room, through a one-way window and over a video monitor. The observer begins a trial when the infant appears ready, records his judgment before the trial ends, and receives feedback at the end of the trial. To encourage the infant to respond to the sound, a mechanical toy in the test booth is activated whenever the observer scores a hit.

All sessions in this study consisted of two phases, a training phase in which the observer was required to achieve a criterion performance with the stimulus at a clearly audible level, and a testing phase in which threshold was estimated. The specifics of these two phases varied with the test condition and are described below. Beside the specific criterion for ending a session in each test condition, sessions were ended if the infant became too fussy or sleepy to continue. If a session was not acceptable for any reason, the entire procedure was repeated in subsequent visits, until a threshold was obtained, the infant's age exceeded criterion, or the parent and infant failed to return for additional sessions.

In all conditions, infant and adult listeners were tested in as similar a way as possible. Adults were told that they were supposed to raise their hands when they "heard the sound that would make the mechanical toy come on." An experimenter in the control room recorded a "yes" response when the adult raised his or her hand. The training and testing procedures were the same for all age groups.

### C. Adaptive methods

#### *1. Subjects*

A total of 34 three-month-olds (average age 2.91 mo, range 2.56–3.21 mo) and 72 six-month-olds (average age 5.92 mo, range 5.54–6.26 mo) were tested. Each subject was screened for middle ear dysfunction using tympanometry at the conclusion of the test session. Eight 3-month-olds and eleven 6-month-olds were excluded because they failed tympanometric screening on the test day. Nine 20- to 30-year-old adults also participated; all passed the tympanometric screen. All subjects were healthy on the test date, with no family history of congenital hearing loss, no personal history of hearing dysfunction, no more than two prior episodes of ear infection, and no ear infection within 3 weeks prior to testing.

#### *2. Test methods*

At the beginning of each session, the level of the stimulus was fixed at 85 dB SPL and the mechanical toy was activated at the conclusion of each signal trial regardless of the observer's judgment. Signal and no-signal trials occurred with equal probability. This training phase continued until the observer achieved four of the last five signal trials correct and four of the last five no-signal trials correct.

Threshold was estimated in the test phase which followed. Two different adaptive algorithms were used to estimate threshold. The first was a modification of the hybrid technique described by Hall (1981). In this technique,

PEST rules (Taylor and Creelman, 1967) are used to adjust stimulus level during the run, but threshold is estimated using a maximum-likelihood criterion on the basis of all the trials completed during the run. Threshold is taken as the 70% correct point on the best fitting psychometric function. The details of the application of Hall's method to infant testing are described by Spetner and Olsho (1990). A modification to Hall's method made here was to present four signal trials at a given level before making a decision to change the intensity. The intensity was increased if the observer got fewer than two of the four signal trials correct, and reduced if the observer got three or four signal trials correct at an intensity. The same intensity was maintained if the observer got two of the four trials correct, then increased on the sixth trial if the observer was incorrect on the fifth trial or decreased if he was correct. The purpose of this modification was to make the staircase more resistant to short lapses of attention. Signal and no-signal trials occurred with equal probability, but only performance on signal trials affected the staircase. The session was considered acceptable only if the observer maintained a false alarm rate below 0.35. This was the method used to estimate thresholds for long duration tones by Olsho *et al.* (1988). Starting level for the staircase was 85 dB SPL; the first step size was 10 dB. Testing continued until at least four reversals and 20 signal trials at no fewer than four stimulus intensities had been obtained.

The second adaptive algorithm was a one-up, two-down rule (Levitt, 1971). Signal and no-signal trials occurred with equal probability. If the observer got two consecutive trials correct, hit or correct rejection, the level was reduced on the next trial. If the observer got one trial incorrect, miss or false alarm, the level was increased on the next trial. Step size was varied according to the PEST rules. The starting level of the staircase was 85 dB SPL and the first step size was 10 dB. The session continued until at least eight reversals had been obtained; threshold was defined as the average of all reversals except the first two.[1] Examples of sessions using the two adaptive procedures are shown in Fig. 1.

A comparison of these two methods may provide an opportunity to evaluate the effects of response criterion on threshold. To simplify discussion, assume that the probability but not the "amplitude" of the infant's response varies with signal level, and that the infant's response does not habituate over trials. Assume further that the observer's general goal is to achieve the lowest threshold possible.

In the modified hybrid procedure, only responses on signal trials affect the value of the threshold. If the acceptable false alarm rate is fairly low, the underlying evidence distributions are normal and equivariate, and the infant/observer team maintains a constant response criterion, it turns out that the false alarm rate will remain approximately constant as sensitivity changes (Hertzog, 1980). Thus a reasonable approach would be to maintain a conservative response criterion that maximizes hit rate, but keeps the false alarm rate at or below 0.35.

In the one-up, two-down procedure, responses on both signal and no-signal trials affect the value of the threshold. In order to get the lowest threshold in this procedure, an observer should try to maximize percent correct. When signal and no-signal trials occur with equal probability, an unbiased observer will maximize percent correct (Green and Swets, 1966). If these scenarios are correct, one would expect that modified hybrid thresholds would be higher than one-up, two-down thresholds.

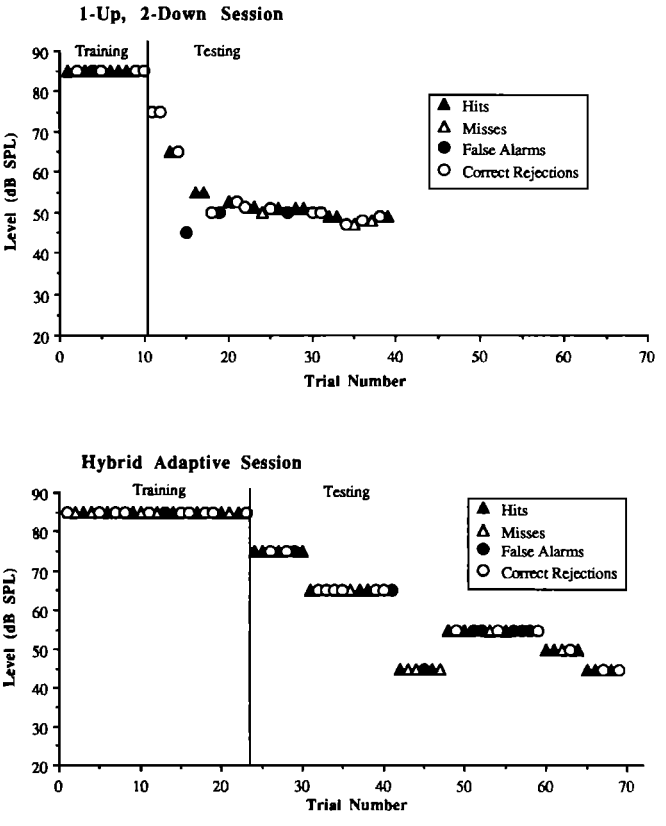


FIG. 1. Examples of trial-by-trial protocols for two adaptive runs, from two individual infants. The top panel shows a hybrid adaptive run for a 6-month-old; the bottom panel shows a one-up, two-down run for a 3-month-old.

These methods also differ in that the underlying psychometric function in the modified hybrid procedure ranges from the false alarm rate to, potentially, 1.00; in the one-up, two-down procedure the psychometric function ranges from 0.50 to, potentially, 1.00. The constraint on the range of the psychometric function in the second procedure should increase variability in threshold estimates (McKee *et al.*, 1985), but only a direct comparison between the two methods can establish the importance of this variability relative to the variability in the behavior of infants.

### D. Method of constant stimuli

#### *1. Subjects*

A total of eighty-two 3-month-olds (average age 3.22 mo, range 2.82–3.51 mo) and forty-six 6-month-olds (average age 6.07 mo, range 5.54–6.49 mo) were tested. Eleven 3-month-olds and fifteen 6-month-olds were excluded because they failed tympanometric screening on the test day. Twenty-two 20- to 30-year-old adults also participated; all passed the tympanometric screen. These subjects met the same inclusion criteria as the subjects tested adaptively.

### 2. Test method

At the beginning of each session, the level of the stimulus was fixed at 85 dB SPL and the mechanical toy was activated at the conclusion of each signal trial regardless of the observer's judgment. The odds of a signal trial were 4 to 1. This initial training phase continued until the observer achieved four of the last five trials correct with the constraint that at least one correct rejection occurred.

A second training phase followed. In this phase, signal and no-signal trials occurred with equal probability, but on signal trials, one of four levels was presented at random. Signal intensities of 85, 76.6, 68.3, and 60 dB SPL were presented to 3-month-olds; intensities of 85, 70.5, 60.5, and 50.5 dB SPL were presented to 6-month-olds. The reason for varying signal intensity was to condition the infant to respond to the stimulus regardless of intensity. In addition, the reinforcer was activated only after the observer scored a hit. This was to teach the infant that reinforcement was contingent upon his or her response. This phase continued until the observer was correct on four of the previous five signal trials and four of the previous five no-signal trials.

During the test phase of the session, signal and no-signal trials were presented with equal probability. On signal trials, one of four levels was presented at random. Levels of 60.0, 53.4, 46.8, and 40.0 dB SPL were presented to 3-month-olds; levels of 50.5, 46.8, 37.6, and 30.5 dB SPL were presented to 6-month-olds, and levels of 25, 18.6, 12.3, and 5 dB SPL were presented to adults. Infant sessions continued as long as the infant's state allowed. Additional sessions were attempted on later dates until the infant completed at least 60 test trials. Adult sessions continued until 240 test trials had been completed, with a break after 120 trials. Acceptable sessions had false alarm rates below 0.25. Threshold was defined as the 70 percent "yes" point on the best fitting psychometric function.[2]

## III. RESULTS

### A. Adaptive procedures versus the method of constant stimuli

Too few hybrid thresholds were obtained to estimate an average for 3-month-olds, so these will not be discussed further. Average one-up, two-down threshold for 3-month-olds was remarkably similar to that obtained using the method of constant stimuli. The same was true of the average thresholds obtained using the two adaptive techniques for 6-month-olds and in the method of constant stimuli. The means and standard deviations are listed in Table I. For the method of constant stimuli, the thresholds are based on the first 30 trials collected for each subject. For both 3- and 6-month-olds, the range of average thresholds across methods is less than 2 dB. Given that the range of the average adult threshold estimates is about 2 dB, the consistency among estimates for the infants is amazingly good. Thus, in terms of an estimate of average sensitivity obtained in OPP, if a threshold is obtained at all, each of these procedures gives the same answer.

As Table I shows, adaptive and method of constant stimuli methods differ in the between-subject variability obtained. Both the hybrid and the one-up, two-down methods produce more variable results than the method of constant stimuli. To some degree this is not a surprising result. If an observer has a string of bad luck or the infant becomes inattentive near the beginning of an adaptive run, the staircase can "level off" at a suprathreshold level. On the other hand, a string of good luck timed properly can drive the staircase to quite low levels. In the method of constant stimuli, a stretch of good luck or bad luck will, on the average, affect all test levels equally. The net effect will be to produce a flat psychometric function, and thus the session will not be included in the final threshold estimate. Thus the opportunity for variable outcomes is greater in the adaptive procedures. At the same time, this difference clearly does not produce a bias in the average threshold.

TABLE I. Mean thresholds, number of thresholds, and standard deviations obtained using three procedures.

| Age | | Hybrid adaptive | One-up, two-down | Method of constant stimuli |
|---|---|---|---|---|
| 3 months | Mean | ··· | 52.43 | 51.22 |
| | s.d. | ··· | 15.58 | 4.59 |
| | *N* | ··· | 11 | 16 |
| 6 months | Mean | 38.71 | 38.82 | 39.19 |
| | s.d. | 14.16 | 13.56 | 5.45 |
| | *N* | 23 | 3 | 21 |
| Adults | Mean | 15.32 | 13.15 | 15.37 |
| | s.d. | 6.34 | 5.90 | 2.59 |
| | *N* | 5 | 4 | 22 |

### B. Stability of threshold estimates in the method of constant stimuli

Eleven 3-month-olds and ten 6-month-olds completed 60 test trials. The method of constant stimuli threshold based on 60 trials was generally the same as that based on 30 trials. Thresholds based on the first 30 trials collected (block 1) were compared to thresholds based on the second 30 trials collected (block 2).[3] The average thresholds are shown in Table II. There was a slight tendency for average infant thresholds to increase between blocks 1 and 2, but this difference is less than 3 dB. A multivariate analysis of variance showed that the neither effect of block [$F(1,17) = 2.05$, $p = 0.17$] nor the block × age interaction [$F(1,17) = 1.34$, $p = 0.26$] were significant.[4] Thus, if habituation occurs, it does not seem to affect the average threshold. The fact that block 2 thresholds were, if anything, a little less variable than block 1 thresholds also argues against any substantial effects of habituation.

If adults were actually improving with practice, finding no change in the infant thresholds with practice would suggest that comparing infants to practiced adults would underestimate their real sensitivity. There was no evidence, however, that the adults improved between 30-trial blocks, for as many as eight such blocks. The average adult threshold never deviated from that in the first 30-trial block by more than

TABLE II. First and second block thresholds and their intercorrelations for three age groups.

| Age | | Block 1 threshold | Block 2 threshold |
|---|---|---|---|
| 3 months | Mean | 51.2 | 51.3 |
| | s.d. | 4.6 | 3.9 |
| | $N$ | 16 | 12 |
| | $r(N,p)$ | −0.79 (10,ns) | |
| 6 months | Mean | 39.2 | 42.15 |
| | s.d. | 5.5 | 4.7 |
| | $N$ | 21 | 10 |
| | $r(N,p)$ | 0.56 (9, <0.05) | |
| Adults | Mean | 15.4 | 14.7 |
| | s.d. | 2.6 | 3.1 |
| | $N$ | 22 | 22 |
| | $r(N,p)$ | 0.81 (22, <0.01) | |

2 dB, with a standard error of the mean less than 1 dB. Individual adults also tended to maintain stable performance across trial blocks; there were no cases where performance improved or worsened substantially with repeated testing. Thus, although even 240 trials represents limited practice in the typical psychophysical experiment, there is no indication that either infants or adults improve in detection performance with practice within this range. Of course, individual adults may have been improving slightly, say 2–3 dB, in threshold; it would have been difficult for us to measure such a small change with the spacing between stimulus intensities used here.

As would be expected, the between-infant variability in threshold tended to be lower when only thresholds based on 60 trials were considered (Table III). At the same time, the number of infants completing 60 trials was smaller than the number completing 30 trials. In fact, although the 3-month-olds' standard error of the mean was a little smaller for 60 trials, the 6-month-olds' standard error was greater for 60 trials compared to 30 trials. Thus, from the standpoint of increasing power in any statistical comparison, increasing the number of trials obtained from each infant actually contributes little.

Finally, we examined the correlation between first and second block thresholds for each age group. The correlations are given in Table II. The correlation between first and second block thresholds was positive and significant for the adults. This result is somewhat surprising in that the variability among normal hearing adult listeners is not generally great enough to support a correlation of this magnitude (e.g., McFadden and Wightman, 1983). Although this may be related to the fact that these listeners were not well trained, the failure to find improvements in performance over additional trial blocks, noted earlier, argues against that interpretation.

The correlations were significant, though not as impressive, for 6-month-old infants. This suggests that 6-month-olds are performing in a relatively consistent manner across trial blocks. In addition, it is clear that other factors that influence threshold in these infants are not stable across trial blocks. In the case of 3-month-olds, in fact, one would have to conclude that there is little stability in performance across trial blocks. None of the correlations were significant and positive.[5,6]

To summarize, neither average infant threshold nor variability in threshold reliably changes between the two 30-trial blocks of trials; this argues against simple habituation, or progressive decrease in responsiveness over the course of a session. At the same time, factors other than auditory sensitivity clearly influence infant threshold estimates, particularly those of 3-month-olds. It is also worth making note of the finding that increasing the number of trials required for a threshold estimate may not have the desired effect of increasing statistical power.

TABLE III. Variability of thresholds estimated in the method of constant stimuli.

| Age | Number of trials | Standard deviation | $N$ | Standard error |
|---|---|---|---|---|
| 3 months | 30 | 4.59 | 16 | 1.15 |
| | 60 | 1.44 | 10 | 0.46 |
| 6 months | 30 | 5.45 | 21 | 1.19 |
| | 60 | 5.15 | 9 | 1.72 |

### C. Effects of response bias on thresholds

#### *1. Adaptive procedures and bias*

Our initial prediction was that a reasonable criterion for an observer in the hybrid adaptive procedure would be that which allowed him to maintain a constant false alarm rate of 0.35 at all levels, while maximizing hit rate. In the one-up, two-down procedure where both hits and correct rejections determine the threshold, the lowest thresholds would be obtained by an unbiased observer. We tested this prediction by examining the hit and false alarm rates for each threshold obtained in an adaptive procedure. In order to avoid averaging hit and false alarm rates over different sensitivities (i.e., different signal levels), only trials falling within 2 dB of the final threshold after the staircase converged on a threshold were counted in this analysis. To get an idea of the general trends and variability in bias, the hits and false alarm rates were plotted in receiver operating characteristic (ROC) space. These plots are shown in Figs. 2 and 3. Isobias contours are also plotted for reference[7] (e.g., Green and Swets, 1966; Hertzog, 1980).

An observer who maintains a constant false alarm rate of about 0.25 (the average rate observed) would produce points that fall between the $\beta = 1.5$ and $\beta = 1.1$ contours. For the hybrid adaptive procedure, we predicted that the data points would fall in this area of ROC space, and, in fact, many of the data points (closed symbols in Fig. 2) do fall there. There are also many data points in this procedure that seem to fall between $\beta = 1.1$ and $\beta = 0.9$, along the negative diagonal, suggesting that at this signal level, the observer has adopted an unbiased criterion, although the false alarm rate is higher than 0.35 near threshold. In order to meet the false

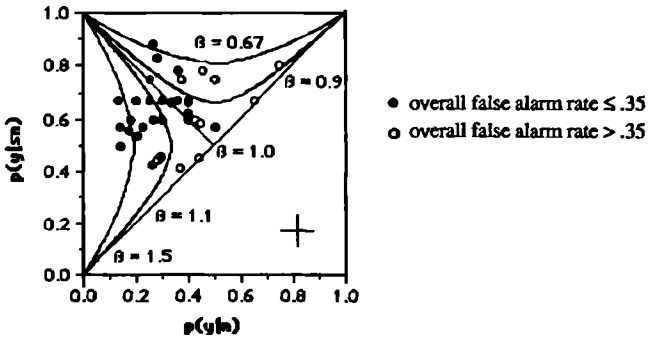


FIG. 2. Receiver operating characteristic plots for 6-month-olds tested with the hybrid adaptive procedure. Each data point represents the hit and false alarm rate of a single infant, on those trials within 2 dB of the estimated threshold. Closed symbols represent the infants included in the final sample. Open symbols represent infants who were excluded from the final sample on account of high false alarm rate. Curves are isobias contours for $\beta$ at the values shown (see text). Bars at lower right represent $\pm$ 1 average binomial standard deviation for these data points.

alarm rate criterion for the session, then, this observer must have a lower false alarm rate at higher signal intensities, which is what would be expected to happen if he remained unbiased throughout the session.

There are, in addition, 12 infants tested in the hybrid procedure whose overall false alarm rate was higher than 0.35, and whose thresholds were not considered in the averages above. The ROC points of these infants near threshold are plotted as the open symbols in Fig. 2. These infants seem to fall into two groups. Half of them fall within the sensitivity range defined by the infants whose overall false alarm rates were under 0.35. The rest are very insensitive, performing around chance levels.

We predicted that the points obtained in the one-up, two-down procedure (Fig. 3) would cluster around the $\beta = 1.0$ contour. Although there are a few outliers, the data points tend to be distributed between $\beta = 1.1$ and $\beta = 0.9$, confirming that prediction.

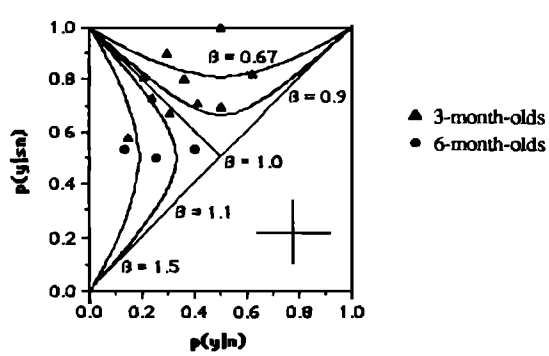


FIG. 3. Receiver operating characteristic plots for 3-month-olds (triangles) and 6-month-olds (circles) tested with the one-up, two-down adaptive procedure. See caption for Fig. 2 and text.

The ROC plots in Figs. 2 and 3 suggest, first, that there is variability between infant/observer teams in response bias, and further, that there are different approaches to maintaining a low false alarm rate in the hybrid technique. How does variability in criterion or in approach affect the estimated threshold? In the case of the one-up, two-down procedure, which was designed to encourage the observer to remain unbiased, there is no evidence that bias affected threshold. The correlation between false alarm rate at threshold and threshold in this procedure was not significant, $r(11) = 0.079$, $p = 0.798$ (footnote 8).

The situation was different for the hybrid procedure, where the correlation between false alarm rate at threshold and threshold, for all of the subjects plotted in Fig. 2, was significant, $r(33) = 0.363$, $p = 0.035$ (footnote 9). If the average threshold for the group is calculated for all 35 infants who had calculable thresholds, the mean is 46.08 dB SPL, as opposed to 38.71 dB SPL for the infants who met the overall false alarm criterion. If all infants ($N = 6$) whose performance was at chance levels at threshold[10] are excluded the mean threshold would be 44.2 dB SPL. Five of these infants were originally excluded on the basis of high false alarm rate. This suggests that if we ensure that the infant/observer team is performing better than would be expected by chance near threshold, then including all infants who produce a threshold, regardless of false alarm rate, would only increase the average threshold estimate by about 5 dB.

### 2. Method of constant stimuli and bias

In the case of the method of constant stimuli, signal levels and no-signal trials were randomly ordered, such that a no-signal trial could not be assigned to any particular signal level (sensitivity). Thus ROC analysis would not be particularly informative with respect to variability in response bias. We examined the relationship between false alarm rate and threshold in the method of constant stimuli. The average thresholds reported above only included infants whose false alarm rate was less than 0.25. For the present analysis, we attempted to calculate thresholds for all infants who were originally excluded on the basis of false alarm rate. Only eight of eighteen 3-month-olds and three of four 6-month-olds originally excluded for false alarm rate had monotonic psychometric functions. As was the case for thresholds obtained using the adaptive procedure, average threshold for infants with high false alarm rates was slightly higher (eight 3-month-olds: $M = 47.1$, s.d. $= 7.1$; three 6-month-olds: $M = 42.3$, s.d. $= 0.76$) than that of the infants who met criterion (Table I). If all infants who only failed the false alarm rate criterion had been included in the means, average threshold would have increased by less than 1 dB.

Is false alarm rate correlated with threshold for individual infants? The analysis of mean differences between infants grouped by false alarm rate suggests that should not be the case. In fact, the correlations were nonsignificant for both 3- and 6-month-olds [$r_{3\,mo}(22) = -0.06$, $p > 0.25$ and $r_{6\,mo}(30) = 0.26$, $p > 0.05$].

Thus, as long as a psychometric function can be fit to an

infant's data, false alarm rate *per se* seems to have only a slight effect on average threshold. Note, further, that only 11 of the 22 sessions not meeting false alarm rate criteria produced monotonic psychometric functions. This is consistent with the finding above that 5 of the 12 infants not meeting false alarm rate criterion in the hybrid adaptive procedure performed at chance levels. On the other hand, only two of thirty-nine 6-month-olds' sessions and one of twenty-six 3-month-olds' sessions meeting all other criteria failed to meet the monotonic psychometric function criterion in the method of constant stimuli. This suggests that excluding sessions on the basis of false alarm rate will exclude most of the sessions during which the infant–observer team shows no evidence of performing at better than chance levels. The false alarm criterion will also exclude some sessions producing estimable thresholds; average thresholds estimated without these sessions, however, will only slightly overestimate the sensitivity of infants as a group.

### D. Other data with respect to threshold estimation procedures

The analyses above describe the characteristics of the thresholds obtained using two adaptive techniques and the method of constant stimuli with infants. Besides these statistical concerns, one might ask whether one procedure is more successful than the others in terms of the number of trials required to estimate a threshold or the number of sessions actually producing a threshold.

It appears, then, that 30 test trials will produce a reasonable threshold estimate for an infant in OPP when the method of constant stimuli is used. Interestingly, the average one-up, two-down adaptive threshold also required about 30 test trials. The hybrid adaptive method, on the other hand, required nearly twice as many test trials ($M = 54.96$) as either of the other methods. This might help to explain why we had such difficulty getting thresholds from infants using this method.

Success rates, or the proportions of sessions producing thresholds, were about the same for adaptive methods and the method of constant stimuli. As shown in Table IV, about 25% of 3-month-olds' adaptive sessions, excluding sessions in which the infant failed the tympanometric screen, produced a threshold, while about 18% of the method of constant stimuli sessions produced thresholds. For 6-month-olds, the rate was higher, but, again, about the same for different procedures, around 50% in each case. In the adaptive procedures this meant that about 40% of all infants tested finally provided thresholds. The method of constant stimuli appeared to be more successful for testing older infants, with 64% of 6-month-old infants, but only 22% of 3-month-old infants providing thresholds.

For the adaptive procedures, a constant sound intensity was used for the training phases, while for the method of constant stimuli, a variable sound intensity was used. Although these differences in training have nothing to do with the adaptive-constant stimuli distinction *per se*, we did compare the success rates and number of trials to training criteria for the two procedures. Whether or not a variable training intensity was used, the average infant producing a threshold in this experiment, whether a 3- or 6-month-old, took about 20 trials to train, the means ranging from 19.3–21.5. There was no indication of a difference between the two training methods for 6-month-olds; about 25% of the sessions ended before training criterion was met for this age group. One interesting difference between the training procedures emerged in 3-month-olds: While the percent of sessions ending before the 3-month-old reached training criterion using the constant training intensity (one-up, two-down) was about 14%, the percent failing to reach training criterion with the variable training intensity was about 52%. If training criterion was met, however, the 3-month-old was almost as likely to produce a threshold in the adaptive procedure (30% of sessions) as in the method of constant stimuli (36.7% of sessions). Thus it appears that a constant training level may be a better training technique for 3-month-olds, even though it makes little difference for 6-month-olds.

In sum, the one-up, two-down procedure and the method of constant stimuli were about the same in terms of the number of thresholds produced per session attempted and the number of trials required to estimate a threshold. The hybrid adaptive procedure required more test trials than either of the other two methods, and thus, must be considered less useful in the testing of infants.

TABLE IV. Numbers of successful and unsuccessful sessions and reasons for exclusion.

| | Method | | |
|---|---|---|---|
| | Adaptive | | |
| | Hybrid | One-up, two-down | Method of constant stimuli |
| 3-month-olds | | | |
| Thresholds | 2 | 11 | 25 |
| Failed tympanometry | 0 | 8 | 11 |
| Not trained | 7 | 6 | 74 |
| Too few test trials | 0 | 22 | 26 |
| False alarm rate | 1 | na | 8 |
| Nonmonotonic function | 1 | na | 9[a] |
| Variable reversals | na | 4 | na |
| Total sessions | 11 | 51 | 153 |
| 6-month-olds | | | |
| Thresholds | 23 | 3 | 37 |
| Failed tympanometry | 8 | 3 | 15 |
| Not trained | 18 | 0 | 20 |
| Too few test trials | 9 | 2 | 14 |
| False alarm rate | 15 | na | 3 |
| Nonmonotonic function | 8 | na | 3[b] |
| Variable reversals | na | 1 | na |
| Total sessions | 81 | 9 | 92 |

[a] Eight of these infants also had a false alarm rate greater than 0.25.
[b] One of these infants also had a false alarm rate greater than 0.25.

## IV. DISCUSSION

Perhaps the most important implication of the findings described here is that the estimate of average infant sensitivity obtained using a behavioral measure, such as OPP, is

little affected by the test method. A 3-month-old's threshold for a short duration, 1-kHz tone burst is about 50 dB SPL, and a 6-month-old's threshold is about 40 dB SPL, whether an adaptive method or the method of constant stimuli is used. Thus, although we cannot argue that the infant's threshold is unaffected by factors other than auditory sensitivity, whatever the threshold is measuring is relatively stable.

There were some differences between procedures, however. The between-infant variability in threshold was considerably lower when the method of constant stimuli was used, even when the number of trials used to estimate a threshold was no greater than that required in the adaptive procedure. Moreover, variability between subjects in response bias were not great enough to significantly affect thresholds measured by this method. In sum, it would appear that the method of constant stimuli is an appropriate method for testing infants. Of course, in order to use this method, the investigator must know approximately where threshold will be. It may be necessary to make an initial estimate of infant sensitivity using an adaptive approach, and then to test hypotheses using thresholds obtained in the method of constant stimuli.

The possible effect of response bias on adaptive thresholds has been a major issue in single-interval psychophysics, but the current results suggest that it may not be as big a problem as has sometimes been suggested. Infant/observer teams tended to be unbiased responders, and response bias generally did not account for threshold variability. The one exception to this result was in the case of the hybrid adaptive procedure, where false alarm rate accounted for about 9% of the variance in infant threshold. The hybrid procedure should probably be avoided for that reason.

Of course, it is preferable to use a bias-free estimate of infant sensitivity, and such measures are becoming more common. Schneider and Trehub (Schneider *et al.*, 1980, 1986; Trehub *et al.*, 1980) have developed a two-alternative, two-choice procedure to test infants, although it is not without its limitations. A few investigators of infant sensitivity have calculated $d'$ (e.g., Trehub *et al.*, 1990; Weir, 1979), but it is not clear that parametric assumptions are justified in this case. Finally, Werner and Gillenwater (1990) have recently used a confidence rating procedure to generate ROC curves from 2- to 5-week-old infants, and such a procedure may ultimately be applicable to older infants as well.

The significant test–retest reliability for 6-month-olds is encouraging. In this age group, moreover, a given infant tends to produce thresholds within 5 dB of each other on repeated tests. This suggests that OPP might be used to assess the sensitivity of individual 6-month-olds. Of course, other reliable behavioral measures (e.g., visual reinforcement audiometry) are already widely used in the clinical assessment of 6-month-olds. That OPP did not prove to be as reliable for individual 3-month-olds suggests, first, that factors other than auditory sensitivity make a major contribution to the between infant variability in threshold, and, second, that OPP is not yet at the point where it could be applied clinically to this population.

A final note is that the differences between infant and adult thresholds for the short-duration tone bursts in this experiment is about 10 dB greater than the difference between infants and adults in threshold for longer duration stimuli (Olsho *et al.*, 1988). Thorpe and Schneider (1987) have reported similar results for 6-month-olds and adults. This means that increasing the duration of a sound improves performance more for an infant than it does for an adult. Since adults are perfect integrators of stimulus energy over durations of as long as 200 ms, this suggests that infants are at a particular disadvantage with short-duration sounds for reasons that are not related to auditory sensitivity. Rather, infants may be less attentive to short duration sounds, or find them less interesting (Gray, 1990). This may also be why we have greater difficulty obtaining thresholds from infants for these sounds as opposed to other stimuli.

## ACKNOWLEDGMENTS

Thanks to Jo Ann Chavira-Bash, Jay Gillenwater, Heather Taylor, and Ross Jurgensen for help in data collection; to Jill Bargones, Pat Feeney, and Lisa Rickard for data collection and critical readings of the manuscript; and to Tino Trahiotis for suggesting the ROC analysis of the adaptive data. This research was supported by NIH Grant No. DC00396 to L. A. Werner.

[1] Levitt (1971) showed that the one-up, two-down rule converges on the 70.7% correct point on the psychometric function in a two-alternative, forced-choice paradigm. We carried out simulations using the current version of the rule to verify that it converged on the same point. A psychometric function with slope and intercept equal to the average established in initial tests was assumed. In addition, the assumption was made that the observer was unbiased, but for each simulation run, a fixed minimum false alarm rate was chosen. One thousand simulated sessions were completed for each run. If a false alarm rate close to zero was assumed, the current one-up, two-down rule converged on the 70% correct point. As the false alarm rate was increased up to 0.35, the rule converged on lower points. However, the effect was to change the average threshold by less than 2 dB. Thus we were confident that the statistical properties of the adaptive rule would have only minor effects on the average infant threshold obtained.

[2] Three different techniques for fitting psychometric functions and estimating thresholds were originally compared. First, psychometric functions were fit to the data points using probit analysis (Finney, 1970). This procedure has the advantage of allowing the estimation of both the upper and lower asymptotes of the psychometric function, thus "correcting" for the infant's tendency to respond in a less than perfect manner. However, recent reports have suggested that probit fits can be inaccurate or misleading (McKee *et al.*, 1985; O'Regan and Humbert, 1989), especially with small numbers of trials. Further, in our own experience with the hybrid adaptive procedure, probit fits occasionally give threshold estimates that seem out of line with the original data.

Consequently, we also estimated thresholds by linear interpolation between the two data points spanning the 70% "yes" point in the raw data. Nonmonotonicities in the data sometimes made it difficult to estimate a threshold in this way. If averaging two adjacent points on the curve eliminated the nonmonotonicity, the curve was "smoothed" in this way before interpolating the threshold point. If this simple smoothing procedure did not produce a monotonic function, no threshold was estimated. While this technique has the advantage of simplicity, it has the disadvantage that infants with high response rates would be expected to produce lower thresholds, regardless of sensitivity.

To attempt to account for differences in response rates, we also used the empirical false alarm rate for the session to rescale the psychometric curve, so that it extended from 0 to 1.0 $p$*("yes"), and then performed a

probit fit to obtain the best psychometric function. The formula for accomplishing this transformation is

$$p^*(\text{"yes"}|s_i) = [p(\text{"yes"}|s_i) - p(\text{"yes"}|n)]/[1 - p(\text{"yes"}|n)],$$

where $p^*(\text{"yes"}|s_i)$ is the rescaled $p(\text{"yes"})$ at signal level $i$, $p(\text{"yes"}|s_i)$ is the original proportion of "yes" responses at that level, and $p(\text{"yes"}|n)$ is the false alarm rate. In essence, this amounts to "correcting for guessing" using the false alarm rate as an estimate of the guessing rate. The transformation is a simple extension of Abbott's rule, as discussed by Finney (1970) and McKee *et al.* (1985).

Each of these three techniques produced essentially the same average threshold, with approximately the same variability. The thresholds reported here are based on linear interpolation, because it was the simplest procedure and actually produced somewhat less variable results than the other techniques. None of the results reported in the method of constant stimuli were substantially different when the other estimation techniques were used.

[3] For the 3-month-olds, six infants completed all 60 trials in one session, while five required two test sessions. For the 6-month-olds, two infants completed 60 trials in one session, eight required two sessions, and one required three sessions.

[4] The main effect of age, incidentally, was highly significant, $F(1,17) = 11.74$, $p = 0.003$.

[5] As noted above, some infants completed 60 trials in one session, while others required more than one session. The correlation analysis is reported for all infants. If one looks at the correlations separately for infants requiring one and two sessions within each age group, the results are not changed substantially. For 6-month-olds only, elimination of the two infants who completed testing in one session did increase the correlation, $r(7) = 0.622$. This suggests that the factors that contribute to variability in infant performance have a more pronounced effect within a single long session, than between two shorter sessions. However, given the small number of subjects involved, this conclusion is necessarily tentative.

[6] Obviously if we had conducted two-tailed significance tests on these correlations, the block 1 versus block 2 correlation for 3-month-olds, for thresholds estimated by linear interpolation, would have been significant, but negative. Since we cannot begin to interpret such a result, we have chosen to ignore it for the present.

[7] $\beta$ is a parametric measure; the underlying distributions of sensory evidence are assumed to be normal and equivariate. While it is not clear that these assumptions are justified for the case of OPP, there are no generally accepted alternative nonparametric bias measures, and these contours are offered to orient the reader rather than as an absolute quantification of bias for these data.

[8] The correlation between $\beta$, a parametric measure of bias, and threshold was also not significant, $r(11) = 0.353$, $p = 0.237$.

[9] The correlation between $\beta$ and threshold was about the same, $r(33) = 0.34$, $p = 0.05$.

[10] Chance was defined as a $d'$ less than 0.2.


Finney, D. J. (1970). *Probit Analysis* (Cambridge U. P., Cambridge).

Gray, L. (1987). "Signal detection analyses of delays in neonates' vocalizations," J. Acoust. Soc. Am. **82**, 1608–1614.

Gray, L. (1990). "Development of temporal integration in newborn chickens," Hear. Res. **45**, 169–178.

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hall, J. L. (1981). "Hybrid adaptive procedure for estimation of psychometric functions," J. Acoust. Soc. Am. **69**, 1763–1769.

Hertzog, C. (1980). "Measurement of response bias in aging research," in *Aging in the 1980s*, edited by L. W. Poon (American Psychological Association, Washington, DC), pp. 568–591.

Levitt, H. (1971). "Transformed up–down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

McFadden, D., and Wightman, F. L. (1983). "Audition: Some relations between normal and pathological hearing," Annu. Rev. Psych. **34**, 95–128.

McKee, S. P., Klein, S. A., and Teller, D. Y. (1985). "Statistical properties of forced-choice psychometric functions: Implications of probit analysis," Percept. Psychophys. **37**, 286–298.

Moore, J. M., Thompson, G., and Thompson, M. (1975). "Auditory localization of infants as a function of reinforcement conditions," J. Speech Hear. Disord. **40**, 29–34.

O'Regan, J. K., and Humbert, R. (1989). "Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used," Percept. Psychophys. **46**, 434–442.

Olsho, L. W., Koch, E. G., Carter, E. A., Halpin, C. F., and Spetner, N. B. (1988). "Pure-tone sensitivity of human infants," J. Acoust. Soc. Am. **84**, 1316–1324.

Olsho, L. W., Koch, E. G., and Halpin, C. F. (1987a). "Level and age effects in infant frequency discrimination," J. Acoust. Soc. Am. **82**, 454–464.

Olsho, L. W., Koch, E. G., Halpin, C. F., and Carter, E. A. (1987b). "An observer-based psychoacoustic procedure for use with young infants," Dev. Psychol. **23**, 627–640.

Schneider, B. A., Trehub, S. E., and Bull, D. (1980). "High-frequency sensitivity in infants," Science **207**, 1003–1004.

Schneider, B. A., Trehub, S. E., Morrongiello, B. A., and Thorpe, L. A. (1986). "Auditory sensitivity in preschool children," J. Acoust. Soc. Am. **79**, 447–452.

Spetner, N. B., and Olsho, L. W. (1990). "Auditory frequency resolution in infancy," Child Dev. **61**, 632–652.

Taylor, M. M., and Creelman, C. D. (1967). "PEST: Efficient estimates on probability functions," J. Acoust. Soc. Am. **41**, 782–787.

Teller, D. Y. (1985). "Psychophysics of infant vision: Definitions and limitations," in *Measurement of Audition and Vision in the First Year of Postnatal Life: A Methodological Overview*, edited by G. Gottlieb and N. Krasnegor (Ablex, Norwood, NJ), pp. 127–145.

Thorpe, L. A., and Schneider, B. A. (1987). "Temporal Integration in Infant Audition," Paper presented to the meeting of the Society for Research in Child Development, Baltimore.

Trehub, S. E., Schneider, B., Thorpe, L. A., and Judge, P. (1990). "Observational measures of auditory sensitivity in infancy," Dev. Psychol. **27**, 40–49.

Trehub, S. E., Schneider, B. A., and Endman, M. (1980). "Developmental changes in infants' sensitivity to octave-band noises," J. Exp. Child Psychol. **29**, 283–293.

Weir, C. (1979). "Auditory frequency sensitivity of human newborns: Some data with improved acoustic and behavioral controls," Percept. Psychophys. **26(4)**, 287–294.

Werner, L. A., and Gillenwater, J. M. (1990). "Pure-tone sensitivity of 2- to 5-week-old infants," Infant Behav. Dev. **13**, 357–377.